Maximum entropy properties of discrete-time first-order stable spline kernel ⋆

Tianshi Chen [a], Tohid Ardeshiri [a], Francesca P. Carli [b], Alessandro Chiuso [c], Lennart Ljung [a], Gianluigi Pillonetto [c]

[a] *Department of Electrical Engineering, Linköping University, Sweden*

[b] *Department of Engineering, University of Cambridge, Cambridge, United Kingdom*

[c] *Department of Information Engineering, University of Padova, Italy*

Abstract

The first order stable spline (SS-1) kernel is used extensively in regularized system identification. In particular, the stable spline estimator models the impulse response as a zero-mean Gaussian process whose covariance is given by the SS-1 kernel. In this paper, we discuss the maximum entropy properties of this prior. In particular, we formulate the exact maximum entropy problem solved by the SS-1 kernel without Gaussian and uniform sampling assumptions. Under general sampling schemes, we also explicitly derive the special structure underlying the SS-1 kernel (e.g. characterizing the tridiagonal nature of its inverse), also giving to it a maximum entropy covariance completion interpretation. Along the way similar maximum entropy properties of the Wiener kernel are also given.

Key words: System identification, regularization method, kernel structure, maximum entropy.

1 Introduction

A core issue of system identification is the design of model estimators able to suitably balance structure complexity and adherence to experimental data. This is also known as the bias-variance problem in statistical literature. Traditionally, this problem is tackled by applying the maximum likelihood/prediction error method (ML/PEM), see e.g., [1], together with model order selection criteria, such as AIC, BIC and cross validation. Recently, a different method has been introduced in [2] and further developed in [3,4,5]; see also the recent survey [6]. Its key idea is to face the bias-variance problem via well-designed and tuned regularization. More specifically, the impulse response $h(t)$ is modeled as a zero-mean Gaussian process $h(t) \sim \mathrm{GP}(0, k(t,s;\alpha))$, where $k(t,s;\alpha)$ is the covariance (kernel) function, and α is the hyper-parameter vector, see e.g., [7]. The key step is to design a suitable kernel structure which reflects our prior knowledge on the system to be identified, e.g., stability. Once $k(t,s;\alpha)$ is determined, α is tuned by maximizing the marginal likelihood, and then the posterior mean of $h(t)$ is returned as the impulse response estimate.

Several kernel structures have been proposed, e.g., the stable spline (SS) kernel in [2] and the diagonal and correlated (DC) kernel in [4], which have shown satisfying performance via extensive simulated case studies. In view of this, it seems interesting to investigate how, beyond the empirical evidence, the use of these regularized approaches can be justified by theoretical arguments. Different perspectives can be taken, e.g. deterministic arguments in favor of SS and DC kernels are developed in [4] while [8] discusses its link to the Brownian Bridge process which suggests the first order stable spline (SS-1) kernel is a natural description for exponentially decaying impulse responses. In this paper, we will instead work within the Bayesian context, discussing the maximum entropy (MaxEnt) properties of the SS-1 kernel.

The MaxEnt approach has been proposed by Jaynes to derive complete statistical prior distributions from incomplete a priori information [9]. Among all distributions that satisfy some constraints, e.g. in terms of the value taken by a few expectations, the MaxEnt criterion chooses the distribution maximizing the entropy. The

⋆ This paper was not presented at any IFAC meeting. Corresponding author T. Chen. Tel. +46-013284726.

Email addresses: tschen@isy.liu.se (Tianshi Chen), tohid@isy.liu.se (Tohid Ardeshiri), fpc23@cam.ac.uk (Francesca P. Carli), chiuso@dei.unipd.it (Alessandro Chiuso), ljung@isy.liu.se (Lennart Ljung), giapi@dei.unipd.it (Gianluigi Pillonetto).

justification underlying this choice is that the MaxEnt distribution, subject to available knowledge, is the one that can be realized in the greatest number of ways, see also Jaynes' Concentration Theorem [9]. A preliminary study on the MaxEnt property of kernels for system identification was developed in [10]. Working in continuous time (CT), the problem was to derive the MaxEnt prior using only information on the smoothness and exponential stability of the impulse response. The arguments in [10] were however quite involved, mainly due to the infinite-dimensional nature of the problem and the fact that the differential entropy rate of a generic CT stochastic process is not well-defined. Another recent contribution is [11] where, under Gaussian and uniform sampling assumptions, it is shown that the SS-1 kernel matrix can be given a MaxEnt covariance completion interpretation [12], that is then exploited to derive its special structure (namely that it admits a tridiagonal inverse with closed form representation as well as factorization).

In this paper, we study the MaxEnt properties of the *discrete-time* (DT) SS-1 kernel. We first formulate the MaxEnt problem solved by the DT SS-1 kernel without Gaussian and uniform sampling assumptions. Then, we extend the result of [11] and link it to our former result: under general sampling assumption, we show that the SS-1 kernel matrix is the solution of a maximum entropy covariance extension problem [12] with band constraints. This results in the well-known tridiagonal structure of the kernel's inverse, which can be also used for efficient numerical implementations [13],[15, Section 5]. As a byproduct, we discuss the MaxEnt properties of the DT Wiener process and its relation with the tridiagonal structure of the inverse of its kernel.

2 MaxEnt property of the Wiener and the SS-1 kernels

Recall that the differential entropy $H(X)$ of a real-valued continuous random variable X is defined as $H(X) = -\int_S p(x)\log p(x)\mathrm{d}x$, where $p(x)$ is the probability density function of X and S is the support set of X.

In the sequel, the objects mainly considered are real-valued DT stochastic processes defined on an ordered index set $\mathcal{T} = \{t_i | 0 \le t_i < t_{i+1}, i = 0, 1, \cdots, \infty\}$.

A real-valued DT stochastic process $w(t)$ with $t \in \mathcal{T}$ is called a white Gaussian noise if $w(t)$ is identically independently Gaussian distributed with mean $\mathbb{E}(w(t)) = 0$ and variance $\mathbb{V}(w(t)) = c$.

2.1 DT Wiener process

The white Gaussian noise has well-known MaxEnt property. On top of it, we can construct a more complex Gaussian process with MaxEnt property which is crucial to derive the MaxEnt property of the SS-1 kernel.

Lemma 1 [1] *Construct a Gaussian process $g(t)$:*

$$\begin{aligned} &g(t_0) = 0 \text{ with } t_0 = 0, \\ &g(t_k) = \sum_{i=1}^{k} w(t_i)\sqrt{t_i - t_{i-1}}, k = 1, 2, \cdots \end{aligned} \tag{1}$$

For any $n \in \mathbb{N}$, it is the solution to the MaxEnt problem

$$\begin{aligned} \underset{h(t)}{\text{maximize}} \quad & H(h(t_1), h(t_2), \cdots, h(t_n)) \\ \textit{subject to} \quad & \mathbb{V}(h(t_i) - h(t_{i-1})) = c\,(t_i - t_{i-1}) \\ & \mathbb{E}(h(t_i)) = 0, i = 1, \cdots, n \end{aligned} \tag{2}$$

where it is assumed that $h(t_0) = 0$ for $t_0 = 0$.

The resulting Gaussian process (1) is actually the DT Wiener process because it satisfies $g(t_0) = 0$, $g(t)$ is Gaussian distributed with zero mean, and has independent increments with $g(t_i) - g(t_j) \sim \mathcal{N}\big(0, c(t_i - t_j)\big)$ for $0 \le t_j < t_i$. It can be verified that the DT Wiener process has zero mean and covariance (kernel) function:

$$\text{Wiener:} \quad K^{\text{Wiener}}(t, s; c) = c\min(t, s),\ t, s \in \mathcal{T} \tag{3}$$

2.2 The first order SS kernel

Based on Lemma 1, we can derive the MaxEnt property for the SS-1 kernel:

$$\begin{aligned} \text{SS-1:} \quad & K^{\text{SS-1}}(t, s; \alpha) = c\min(e^{-\beta t}, e^{-\beta s}), \\ & \alpha = [c\ \beta]^T, c \ge 0, \beta > 0,\ t, s \in \mathcal{T} \end{aligned} \tag{4}$$

It is also introduced independently in a deterministic argument in [4] and called the tuned correlated (TC) kernel. It is fair to call (4) the SS-1 kernel here, since the "stable" time transformation involved in deriving the SS-1 kernel plays a key role in the following theorem.

Theorem 1 *Let $w(\cdot)$ be a white Gaussian noise with mean zero and variance c. Then the stochastic process*

$$\begin{aligned} h^o(t_k) = \sum_{i=k}^{n-1} w(e^{-\beta t_i})\sqrt{e^{-\beta t_i} - e^{-\beta t_{i+1}}}, \\ k = 0, \cdots, n-1, h^o(t_n) = 0 \text{ with } t_n = \infty \end{aligned} \tag{5}$$

is a Gaussian process with zero mean and the SS-1 kernel (4) as its covariance function, and for any $n \in \mathbb{N}$, it is the solution to the MaxEnt problem

$$\begin{aligned} \underset{h(t)}{\text{maximize}} \quad & H(h(t_0), h(t_1), \cdots, h(t_{n-1})) \\ \textit{subject to} \quad & \mathbb{V}(h(t_{i+1}) - h(t_i)) = c\left(e^{-\beta t_i} - e^{-\beta t_{i+1}}\right) \\ & \mathbb{E}(h(t_i)) = 0, i = 0, \cdots, n-1 \end{aligned} \tag{6}$$

[1] All proofs can be found in the Appendix.

where it is assumed that $h(t_n) = 0$ with $t_n = \infty$.

Remark 1 *In the optimization criteria (2) and (6), if we divide the entropy of the sequence of the stochastic process by n and let n go to ∞, then the limit (if exists) becomes the differential entropy rate of the stochastic process [14]. However, the limit does not exist for Gaussian processes (1) and (5), which is the reason why the entropy of a sequence of stochastic processes is used here instead.*

3 Special structure of Wiener and SS-1 kernels and their MaxEnt interpretation

In what follows, we let $c = 1$ and consider kernel matrix P with dimension $n \geq 3$ defined as

$$P_{i,j} = K(t_i, t_j; \alpha),\ i, j = 1, \cdots, n,\ t_i, t_j \in \mathcal{T} \tag{7}$$

where $P_{i,j}$ denotes the (i,j)th element of the matrix P and K is either the Wiener kernel (3) or the SS-1 kernel (4). We find that P has some special structure, e.g., its inverse is tridiagonal and its square root has closed-form expression. These special structure can be used to improve the stability and efficiency of the implementation solving the marginal likelihood maximization, see e.g., [13, Remark 4.2], [15, Section 5].

Proposition 1 *Consider the Wiener kernel (3) and the SS-1 kernel (4). Then the following results hold:*

(a) for the Wiener kernel, $\det(P^{Wiener}) = t_1 \Pi_{k=1}^{n-1}(t_{k+1} - t_k)$ *and* $(P^{Wiener})^{-1}_{i,j}$ *is equal to*

$$\begin{cases} \frac{t_2}{t_1(t_2-t_1)}, & i = j = 1, \\ \frac{t_{i+1}-t_{i-1}}{(t_{i+1}-t_i)(t_i-t_{i-1})}, & i = j = 2, \cdots, n-1, \\ \frac{1}{t_n - t_{n-1}}, & i = j = n, \\ 0, & |i-j| > 1 \\ -\frac{1}{\max(t_i,t_j) - \min(t_i,t_j)}, & \text{otherwise}, \end{cases}$$

(b) for the SS-1 kernel, $\det(P^{SS\text{-}1}) = e^{-\beta t_n} \Pi_{k=1}^{n-1}(e^{-\beta t_k} - e^{-\beta t_{k+1}})$ *and* $(P^{SS\text{-}1})^{-1}_{i,j}$ *is equal to*

$$\begin{cases} \frac{1}{e^{-\beta t_1} - e^{-\beta t_2}}, & i = j = 1, \\ \frac{e^{-\beta t_{i-1}} - e^{-\beta t_{i+1}}}{(e^{-\beta t_{i-1}} - e^{-\beta t_i})(e^{-\beta t_i} - e^{-\beta t_{i+1}})}, & i = j = 2, \cdots, n-1, \\ \frac{e^{-\beta(t_{n-1} - t_n)}}{e^{-\beta t_{n-1}} - e^{-\beta t_n}}, & i = j = n, \\ 0, & |i-j| > 1 \\ -\frac{1}{e^{-\beta \min\{t_i,t_j\}} - e^{-\beta \max\{t_i,t_j\}}}, & \text{otherwise}, \end{cases}$$

Corollary 1 *Consider the Wiener kernel (3) and the SS-1 kernel (4). Then the following results hold:*

(a) for the Wiener kernel,

$$(P^{Wiener})^{-1} = W^T W \tag{8}$$

where W is upper bidiagonal with

$$W(i,i) = -\frac{t_{i+1}}{t_i} W(i, i+1) = \sqrt{\frac{t_{i+1}}{t_i} \frac{1}{t_{i+1} - t_i}},$$
$$i = 1, \cdots, n-1,\ W(n,n) = \sqrt{1/t_n}$$

(b) for the SS-1 kernel,

$$(P^{SS\text{-}1})^{-1} = S^T S \tag{9}$$

where S is upper bidiagonal with

$$S(i,i) = -S(i, i+1) = \frac{1}{\sqrt{e^{-\beta t_i} - e^{-\beta t_{i+1}}}},$$
$$i = 1, \cdots, n-1,\ S(n,n) = \sqrt{\frac{e^{\beta(t_n - t_{n-1})} - 1}{e^{-\beta t_{n-1}} - e^{-\beta t_n}}}$$

Remark 2 *From (8) and (9), decomposing $P = UU^T$ for upper triangular U has closed form expression. For the Wiener kernel, $U = W^{-1}$ with $U_{i,j} = (W_{i,i})^{-1} t_i / t_j$ for $i \geq j$, $i, j = 1, \cdots, n$. For the SS-1 kernel, $U = S^{-1}$ with $U_{i,j} = (S_{i,i})^{-1}$ for $i \geq j$, $i, j = 1, \cdots, n$.*

Remark 3 *Recall from e.g., [12] that if $X \sim \mathcal{N}(0, P)$ with $P^{-1}_{i,j} = 0$, then X_i and X_j are conditionally independent given X_k with $k \neq i, j$ where X_k is the kth element of X. This means that the Wiener and SS-1 kernels correspond to sparse representation, see e.g., [12] for details and also the proof of Corollary 1.*

3.1 MaxEnt covariance completion

The fact that the kernel matrices of the Wiener and SS-1 kernels have tridiagonal inverse can be given a MaxEnt covariance completion interpretation.

Recall that a real symmetric matrix A with dimension $n > m+1$ is called an $m-$*band* matrix if $A_{i,j} = 0$ for $|i-j| > m$, and the matrix M is called an extension of A if $M_{i,j} = A_{i,j}$ for $|i-j| \leq m$. Moreover, M is called a positive extension of A if M is positive definite. A positive extension M of the $m-$band matrix A is called a band-extension of A if M^{-1} is an $m-$band matrix.

Theorem 2 *Define $A \in \mathbb{R}^{n \times n}$ as follows:*

$$A_{i,j} = \begin{cases} P^{Wiener}_{i,j}\ (resp.\ P^{SS\text{-}1}_{i,j}),\ |i-j| \leq 1 \\ 0 \qquad\qquad\qquad\qquad |i-j| > 1 \end{cases} \tag{10}$$

Then P^{Wiener} (resp. $P^{SS\text{-}1}$) is the unique band extension of A, and the Gaussian random vector with zero mean and covariance P^{Wiener} (resp. $P^{SS\text{-}1}$) is the unique solution to the MaxEnt covariance completion problem

$$\begin{aligned} &\underset{P}{\text{maximize}} \quad H(X) \\ &\textit{subject to} \quad P \textit{ is any positive extension of } A \end{aligned} \tag{11}$$

where X is a zero mean random vector with covariance matrix P.

Remark 4 *To our best knowledge, for the Wiener kernel (3) the special structure and its MaxEnt interpretation has not been pointed out before. For the SS-1 kernel (4), the result under the uniform sampling assumption is given in [11] and thus is a special case of this paper.*

4 Conclusion

We have shown that a zero mean Gaussian process with the first-order stable spline kernel solves a maximum entropy problem with the constraint that the variance of neighboring impulse response coefficients at $t_i < t_{i+1}$ is proportional to $e^{-\beta t_i} - e^{-\beta t_{i+1}}$, which decays to zero ultimately. Its kernel matrix (also true for the Wiener kernel) solves a maximum entropy covariance completion problem and has special structure, e.g., its inverse is tridiagonal, under general sampling assumptions. Finally, one may wonder if the other kernels, e.g., the diagonal correlated kernel, can be given similar maximum entropy interpretation. The answer is more involved and will be discussed separately, see e.g., [15].

References

[1] L. Ljung. *System Identification - Theory for the User*. Prentice-Hall, Upper Saddle River, N.J., 2nd edition, 1999.

[2] G. Pillonetto and G. De Nicolao. A new kernel-based approach for linear system identification. *Automatica*, 46(1):81–93, 2010.

[3] G. Pillonetto, A. Chiuso, and G. De Nicolao. Prediction error identification of linear systems: a nonparametric Gaussian regression approach. *Automatica*, 47(2):291–305, 2011.

[4] T. Chen, H. Ohlsson, and L. Ljung. On the estimation of transfer functions, regularizations and Gaussian processes - Revisited. *Automatica*, 48:1525–1535, 2012.

[5] T. Chen, M. S. Andersen, L. Ljung, A. Chiuso, and G. Pillonetto. System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques. *IEEE Transactions on Automatic Control*, (11):2933–2945, 2014.

[6] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung. Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, 50(3):657–682, 2014.

[7] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

[8] A. Chiuso, T. Chen, L. Ljung, and G. Pillonetto. On the design of multiple kernels for nonparametric linear system identification. In *Proceedings of the IEEE Conference on Decision and Control*, Los Angeles, CA., 2014.

[9] E. T Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, 1982.

[10] G. Pillonetto and G. De Nicolao. Kernel selection in linear system identification. Part I: A Gaussian process perspective. In *Proc. 50th IEEE Conference on Decision and Control*, pages 4318–4325, Orlando, Florida, 2011.

[11] F. P. Carli. On the maximum entropy property of the first-order stable spline kernel and its implications. In *IEEE Multi-Conference on Systems and Control*, pages 409–414, Nice/Antibes, France, 2014.

[12] A. P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

[13] T. Chen and L. Ljung. Implementation of algorithms for tuning parameters in regularized least squares problems in system identification. *Automatica*, 49:2213–2220, 2013.

[14] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[15] F. P. Carli, T. Chen, and L. Ljung. Maximum entropy kernels for system identification. *arXiv:1411.5620*, 2014.

[16] I. Gohberg, S. Goldberg, and A. Kaashoek. *Classes of Linear Operators*. Operator theory, advances and applications. Birkhäuser Verlag, 1993.

Appendix

Proof of Lemma 1

First, we recall the well-known MaxEnt property of white Gaussian noise.

Lemma 2 *[14, Burg's MaxEnt Theorem, page 417] Consider the white Gaussian noise $w(t)$. For any $n \in \mathbb{N}$, it is the solution to the MaxEnt problem:*

$$\begin{aligned} &\underset{r(t)}{\text{maximize}} \quad H(r(t_0), r(t_1), \cdots, r(t_{n-1})) \\ &\textit{subject to} \quad \mathbb{E}(r(t_i)) = 0, \mathbb{V}(r(t_i)) = c, i = 0, ..., n-1 \end{aligned} \tag{12}$$

Then from (2), define $v(t_i) = \frac{h(t_i) - h(t_{i-1})}{\sqrt{t_i - t_{i-1}}}$, $i = 1, \cdots, n$. We have $\mathbb{E}(v(t_i)) = 0, \mathbb{V}(v(t_i)) = c$, $i = 1, \cdots, n$, and

$$h(t_k) = \sum_{i=1}^{k} v(t_i)\sqrt{t_i - t_{i-1}}, k = 1, 2, \cdots, n \tag{13}$$

Now let $L = [h(t_1)\ h(t_2)\ \cdots\ h(t_n)]^T$, $V = [v(t_1)\ v(t_2)\ \cdots\ v(t_n)]^T$, and B be a lower-triangular matrix with $B_{i,j} = \sqrt{t_j - t_{j-1}}$ for $i \geq j$. Then we have $L = BV$. Apparently, B is nonsingular in that all main diagonal elements are strictly positive. Further noting the property (see e.g., [14, Corollary to Theorem 8.6.4]) that

$H(L) = H(V) + \log\det(B)$ yields that the MaxEnt problem (2) is equivalent to

$$\begin{aligned} \underset{v(t)}{\text{maximize}} \quad & H(v(t_1), v(t_2), \cdots, v(t_n)) + \log\det(B) \\ \text{subject to} \quad & \mathbb{E}(v(t_i)) = 0, \mathbb{V}(v(t_i)) = c, i = 1, \cdots, n \end{aligned} \tag{14}$$

Since the matrix B is independent of $v(t)$ or $h(t)$, the maximum entropy problem (14) is further equivalent to (12). As a result, the optimal $v(t)$ to (14) is the white Gaussian noise $w(t)$. Finally, comparing (13) with (1) yields that the constructed Gaussian process $g(t)$ in (1) is indeed the optimal solution to (2).

Proof of Theorem 1

We first introduce a time transformation, and define

$$\tau_i = e^{-\beta t_{n-i}}, \tag{15}$$

$$f(\tau_i) = h(-\log(\tau_i)/\beta),\ i = 0, \cdots, n. \tag{16}$$

Then the MaxEnt problem (6) is equivalent to

$$\begin{aligned} \underset{f(\tau)}{\text{maximize}} \quad & H(f(\tau_1), f(\tau_2), \cdots, f(\tau_n)) \\ \text{subject to} \quad & \mathbb{V}(f(\tau_i) - f(\tau_{i-1})) = c\,(\tau_i - \tau_{i-1}) \\ & \mathbb{E}(f(\tau_i)) = 0, i = 1, \cdots, n \end{aligned} \tag{17}$$

where it is assumed that $f(\tau_0) = 0$ with $\tau_0 = 0$. By Lemma 1, the optimal solution to (17) is the Gaussian process $g(\tau)$ defined as follows:

$$\begin{aligned} & g(\tau_0) = 0 \text{ with } \tau_0 = 0, \\ & g(\tau_k) = \sum_{i=1}^{k} w(\tau_i)\sqrt{\tau_i - \tau_{i-1}}, k = 1, 2, \cdots \end{aligned} \tag{18}$$

where $w(\tau)$ is the white Gaussian noise defined on $\{\tau_0, \tau_1, \cdots\}$. Finally, noting (16) and (15) yields that the optimal solution to (6) is (5). Apparently, (5) is a Gaussian process with zero mean and the SS-1 kernel as its covariance function. This completes the proof.

Proof of Proposition 1

For the proof of the results hereafter, we only give the proof for the SS-1 kernel and that for the Wiener kernel can be derived in the same way and thus is omitted.

From (5), define $x = [x_1, \cdots, x_n]^T$ with $x_k = h^o(t_k) - h(t_{k-1})$, $k = 1, \cdots, n$. Then we have $x \sim \mathcal{N}(0, Q)$ where Q is a diagonal matrix with $Q_{i,i} = e^{-\beta t_{i-1}} - e^{-\beta t_i}$, $i = 1, \cdots, n$. Moreover, $(P^{\text{SS-1}})^{-1} = V^{-T} Q^{-1} V$ where V is an upper bidiagonal matrix with all main diagonal elements equal to -1 and the first upper off-diagonal elements equal to 1. Apparently, $(P^{\text{SS-1}})^{-1}$ takes the form in part b), which completes the proof.

Proof of Corollary 1

By completing the squares, $\theta^T (P^{\text{SS-1}})^{-1} \theta = \sum_{k=1}^{n-1} S_{k,k}^2 (\theta_k - \theta_{k+1})^2 + S_{n,n}^2 \theta_n^2$ where $\theta \in \mathbb{R}^n$ and θ_k is the kth element of θ. Then (9) follows immediately.

Proof of Theorem 2

We first recall a lemma from band matrix extension problems, that is a result of [16, Theorem 2.1, page 898, Theorem 2.2, page 899, Corollary 1.5, page 945].

Lemma 3 *[16] Assume that A is an m-band matrix with dimension $n > m + 1$ and that the submatrices $[A]_i^{m+i}$, $i = 1, \cdots, n - m$, are positive definite, where $[A]_s^l$ with $s \le l$ denotes the submatrix of A from the sth row (resp. column) to the lth row (resp. column). Then we have:*

(a) M is the unique band extension of A.

(b) The Gaussian random vector with zero mean and covariance matrix M is the unique solution to the MaxEnt problem

$$\begin{aligned} \underset{X}{\text{maximize}} \quad & H(X) \\ \textit{subject to} \quad & P \textit{ is any positive extension of } A \end{aligned} \tag{19}$$

where X is a zero mean random vector with covariance matrix P.

Apparently, A in (10) is a $1-$ band matrix and $[A]_i^{i+1}$, $i = 1, \cdots, n-1$, are positive definite. This means that the results of Lemma 3 hold for A in (10) and the remaining task is to show $M = P^{\text{SS-1}}$, i.e., the optimal solution P^{Opt} of (19) is $P^{\text{Opt}} = P^{\text{SS-1}}$. This task can be accomplished by noting the relation between the problems (19) and (6). Note that the Gaussian process (5) solves the problem (6) and has the SS-1 kernel as its covariance function. Assume $Z \sim \mathcal{N}(0, P)$. Then for $n \ge 3$, the covariance matrix $P^{\text{SS-1}}$ is the optimal solution to

$$\begin{aligned} \underset{P}{\text{maximize}} \quad & H(Z) \\ \text{subject to} \quad & P_{i,i} + P_{i+1,i+1} - 2P_{i,i+1} \\ & = e^{-\beta t_i} - e^{-\beta t_{i+1}}, i = 1, \cdots, n-1 \\ & P \text{ is positive definite} \end{aligned} \tag{20}$$

Also note that the constraint in (19) is a subset of the constraint in (20), hence $H(X) \le H(Z)$ with $X \sim \mathcal{N}(0, P^{\text{Opt}})$ and $Z \sim \mathcal{N}(0, P^{\text{SS-1}})$. Finally, noting that $P^{\text{SS-1}}$ is a positive extension of A and the uniqueness of P^{Opt} yields $P^{\text{Opt}} = P^{\text{SS-1}}$. This completes the proof.